CHAPTER

7

Application of Ultra-High Throughput Sequencing and Microarray Technologies in Pharmacogenomics Testing

Gary Hardiman

Department of Medicine, and BIOGEM, School of Medicine, University of California, San Diego, CA; Computational Science Research Center and Biomedical Informatics Research Center, San Diego State University, San Diego, CA

OUTLINE

DOI: 10.1016/B978-0-12-385467-4.00007-5

INTRODUCTION

The sequencing of the human genome over a decade ago was a critical scientific milestone. The draft human genome sequence has revolutionized the pharmaceutical industry, providing a framework for the identification of novel drug targets, and elucidation of the genetic factors that affect drug metabolism and toxicity and of those that contribute to the wide variability in pharmacological treatment responses. The emergence and wide acceptance of genomic technologies, including microarray and deep sequencing technologies, has allowed geneticists, biologists and pharmacologists to bridge the gap between gene sequence and function on a scale that was not possible previously. These newer technological approaches have been integrated into multiple aspects of the drug discovery process, including target validation, pharmacokinetics and toxicology, and clinical pharmacogenomics [1].

The adoption of a novel technology is met by a paradigm shift in how biological assays are designed and executed. Throughput is increased by at least an order of magnitude and accompanied by an exponential cost reduction compared to older traditional approaches. The economic benefit and efficacy of nascent technologies is often realized by process-miniaturization combined with the multiplexing of millions of reactions. A classic example of this scheme is the DNA microarray which was developed in the early 1990s [2, 3]. DNA arrays or (bio)chips evolved steadily over time from archetypal in-house boutique efforts to robust commercial arrays. This progression was accompanied by the availability of higher density chips with increased content, lower per-sample costs, and concomitant increases in sensitivity, accuracy and precision [3–7]. Frequently, an appreciation of the benefits and long-term impact of nascent technologies is not immediate, as their initial use is restricted to developers, beta-testers and early adopters. The impact of a given technology and its long-term value is not realized until its wider acceptance by the scientific community.

In the past five years, ultra-high throughput or deep sequencing of DNA has transitioned from development to widespread use, with several well-established academic and commercial efforts in place [8–10]. Deep sequencing has reduced the cost per base of DNA sequencing by several orders of magnitude, and has effectively introduced genome sequencing center capability into every laboratory [8]. Incorporation of next-generation sequencing approaches into drug discovery programs has become an important issue for many biotechnology and pharmaceutical companies. As the technology is in a state of flux with constant improvements and upgrades, and with new players being added to this space almost monthly, this creates complexity as to the choice of the most appropriate platform.

Many changes can be expected in the next few years, mirroring the early days in the development of microarray technology, where early adopters were forced to build on their initial investment and commit additional resources in the form of equipment upgrades or hardware replacement to remain current with the technology. Cost, ease of use, versatility, peer review in the form of published data, platform stability and the quality of long-term technical support will continue to guide the technology selection process. Data management also poses challenges, with the need for high-speed fiber-optic networks, and ample storage capacity for long-term data storage, annotation, query and retrieval. The microarray comparison is again appropriate when one considers the current state of analytical tools for next-generation sequencing analysis. Better analytical tools will emerge over time, likely from open source efforts, permitting additional analyses and enhanced information mining from raw data sets compared to the tool kits provided with the instruments themselves.

DNA MICROARRAY

A DNA microarray (commonly referred to as gene or genome chip, cDNA array, DNA chip or biochip) is a collection of DNA features attached to a solid support – commonly silanized glass, plastic, film or silicon. The array features or "spots" contain individual DNA probes which are used to interrogate mRNAs (expression) or polymorphisms (genotype). Most of the arrays in use today contain hundreds to thousands of probes. The utility of this technology is that it permits highly parallel measurements from an individual sample or patient specimen.

In the case of gene expression profiling, the substantial number of data points obtained from a single experiment provides an insight into the state of a transcriptome in, for example, healthy and diseased cells, or cells before and after exposure to a therapeutic. The knowledge obtained from such comparisons permits the identification of gene families and pathways pertinent to a malady or drug treatment, in addition to those that remain unaffected. Similar expression profiles may infer that genes are co-regulated; enabling the formulation of hypotheses about genes with hitherto unknown functions by comparison of their expression patterns to well-characterized genes [11–15]. This approach permits the discovery of DNA biomarkers and facilitates the development of diagnostic and prognostic tests.

The applicability of microarrays in genomics research has expanded with the evolution and maturation of the technology. Biochips have found utility in exon-based gene expression analyses, genotyping and re-sequencing applications, comparative genomic hybridization studies and genome-wide (epigenetic) localization. Biochips are being widely applied to improve the processes of disease diagnosis, pharmacogenomics, and toxicogenomics [12–15].

PHARMACOGENETIC TESTING AND HEALTHCARE

Heterogeneity is observed in the manner in which individuals respond to medications. Clinical observations of inherited differences in drug effects were first noted in the 1950s. By the 1990s, it was well established that inherited differences in drug metabolism and disposition, and genetic polymorphisms in the targets of drug therapy, could have a profound effect on the efficacy and toxicity of medications [16, 17]. Pharmacogenetics is concerned with the relationship between a patient's inherited genetic makeup and his or her response to pharmaceutical drugs. Pharmacogenetic testing aims at determining the underlying genotypic and phenotypic differences in the pharmacodynamics and pharmacokinetics of drug metabolism. Whereas pharmacogenetics refers to genetic differences (variation) in drug metabolism and response, pharmacogenomics refers to the study of the multiplicity of genes that ultimately determine drug behavior. Pharmacogenomics is in essence a whole-genome application of pharmacogenetics, correlating gene expression or single nucleotide polymorphisms (SNPs) with drug efficacy and toxicity. Genetic variability in drug response occurs as a result of molecular alterations in the enzymes involved in the metabolism of a particular drug, in addition to the drug receptors and transport proteins [18].

The grand promise of pharmacogenomics is the development of therapeutics targeted for specific patient subgroups. The vision has been the application of high-throughput molecular diagnostics approaches, DNA microarrays and next-generation sequencing technologies as sensitive screening tools for genetic predisposition to the adverse effects of therapeutics. These approaches would facilitate patient stratification and robust selection of medications and dosages tailored to address inter-individual variability [19, 20]. An advance and fundamental shift in healthcare has been the emergence of personalized medicine [21]. This model emphasizes the customization of healthcare, with all decisions and practices being tailored to individual patients. This approach makes use of genetic and/or other information about individual patients to select or optimize their preventative and therapeutic care.

Drug−drug interactions (DDIs) can have serious consequences such as adverse drug reactions (ADRs), and extreme outcomes including death. A drug interaction occurs when a substance affects the activity of a drug, either increasing or decreasing its efficacy, or, alternatively, a new effect is observed that is not observed with just the drug alone [22−24]. DDIs have become a serious issue, particularly in the care of elderly patients, who are often prescribed a wide variety of medications [25]. ADRs are presently the fourth leading cause of death in US. The economics of drug-related morbidity and mortality has become a pressing issue, with current costs exceeding US$177 billion annually in the USA alone [23].

Pharmacogenetic approaches are being used more widely for therapeutic monitoring and health management of patients, with patient genotyping and stratification performed well in advance of drug treatments, thereby eliminating completely or greatly reducing adverse effects. The testing itself can generally be performed in a non-invasive manner using DNA obtained from saliva, hair root or buccal swab samples, and can provide predictive values for many drugs rather than a single drug [18].

DNA MICROARRAY PLATFORMS

The platforms most widely utilized in the past decade for expression profiling include those developed commercially by Affymetrix, Agilent, Nimblegen, Applied Microarrays CodeLink, and Illumina. Affymetrix (Santa Clara, CA) pioneered this crowded field by developing a GeneChip™ comprising short 25-mer oligonucleotide probes that were fabricated *in situ* using a combination of photolithographic techniques (borrowed from the silicon chip industry) and solid phase DNA synthesis [5, 6]. Agilent (Palo Alto, CA) coupled inkjet printing (developed by Hewlett Packard) with standard phosphoramidite chemistry [7, 8] to synthesize 60-mer probes. Nimblegen (Madison, WI) coupled photolithography and solid phase DNA synthesis but, unlike the related Affymetrix technology, Nimblegen disposed of solid chromium mask photolithography in favor of digital micro-mirrors (DMDs or DLPs). DMDs flip mirrors on and off, thereby providing the highly specific light patterns required for photo-activation and DNA chain extension components of DNA chip synthesis [26]. Nimblegen designed isothermal probes to minimize hybridization artifacts and bias, with probe lengths ranging from 45 to 85 mers, depending on the particular application. Applied Microarrays (Tempe, AZ) utilized a non-contact, piezoelectric dispensing method optimized by Motorola to deposit 30-mer oligonucleotides on a three-dimensional polyacrylamide gel (CodeLink™) matrix [27]. A bead-based microarray technology was developed by Illumina (San Diego, CA) permitting multiplexing of up to 96 samples. These substrates contain thousands of tiny etched wells, into which thousands to hundreds of thousands of 3-micron beads are randomly self-assembled. Gene-specific probes (50 mers) concatenated with “address or zip-code” sequences are immobilized on the bead surface. Once bead assembly has occurred the array is “decoded”, to uncover which bead type containing a particular sequence is present in each well of the substrate [28–30].

The net result of a microarray experiment is a list of significant probe sets or genes that are differentially expressed across given experimental conditions. The standard experimental design utilizes biological replicates, permitting an estimate of the statistical relevance of a given fold-change between a control and treated condition. The subsequent steps in interpretation of this data are appending biological knowledge to this list. This process has matured considerably over the past decade, and relied on the expansion of public databases such as Entrez Gene, Unigene, UniProt, Gene Ontology (GO) and KEGG pathways in addition to commercial proprietary efforts.

In an effort to minimize the need to replicate microarray experiments, and to provide exact descriptions for experimenters wishing to utilize public datasets, gene expression data is today stored in public repositories such as Gene Expression Omnibus (GEO) and EBI Array Express, which have become the major portals for deposition and retrieval of genomics data. The “Minimum Information about a Microarray Experiment (MIAME)” standard today requires experimenters to report in detail their experimental design, sample description, labeling and hybridization protocol, data-imaging conditions and data analysis using a pre-determined set of criteria. In order to publish microarray data in the majority of peer-reviewed journals today, mandatory submission of the dataset to one of the databases precedes publication.

In order to assess the performance and reliability of different microarray platforms, a comprehensive study known as the MicroArray Quality Control (MAQC) project was undertaken. This global effort had FDA (Federal Drug Administration of the United Sates Government) oversight and brought together academic and industrial scientists. The outcome was a detailed assessment of the strengths and weaknesses of the major platforms in terms of their performance and reliability, and better strategies for integrating data from different platforms.

Array comparative genomic hybridization (a-CGH) has today superseded traditional chromosome-based methods for detection of genomic copy number variations. It permits higher resolution levels than chromosome-based methods, facilitating identification of chromosomal changes such as micro-deletions and duplications. The Agilent Human Genome CGH Microarray is one such high-resolution platform that has been widely applied for profiling genome-wide DNA variation without genome amplification or complexity reduction approaches. The output of a CGH experiment is a series of copy number variations

(CNV), namely DNA segments which range in size from under 1 kilobase to several megabases of DNA.

MICROARRAYS AND GENOTYPE

Single nucleotide polymorphisms (SNPs) are highly abundant, with over 10 million present in the human genome. SNPs serve as valuable markers of genome-wide variation. A chromosome region may contain many SNPs, but just a few "tag" SNPs is all that is required to provide information on the pattern of genetic variation. The high costs associated with most SNP detection strategies have until recently made genome-wide approaches impractical. The relatively low-cost and high-throughput capabilities of Illumina bead-based technology made genome-wide approaches a reality. Genome-wide genotyping of defined sets of up to 1 million SNPs can be performed using the Infinium assay, developed by Illumina. A whole-genome amplification step is initially employed to enrich the target DNA up to 1000-fold. Once amplified, the DNA is subsequently fragmented and mobilized by hybridization to SNP-specific primers present on the array. An oligonucleotide primer is hybridized adjacent to the SNP site and is extended with a single labeled dideoxynucleotide terminator corresponding to the minor or major allele. Genotyping calls can then be made, based on the dye-labeled terminator that is incorporated [29–30].

Since their development in 2005, the Illumina whole-genome genotyping (WGGT) arrays have become an important tool for discovering variants that contribute to human diseases and phenotypes. Illumina arrays permit two different types of study: genome-wide association studies (GWAS) and copy-number variant (CNV) analyses. The genome-wide association study (GWA study or GWAS) is an unbiased examination of the entire genome of different individuals to see if any variant is associated with a phenotypic (disease) trait. Single nucleotide polymorphisms (SNPs) are investigated. These studies typically compare the DNA of two groups of participants: people with the disease (cases), and age- and sex-matched people without (controls). If genetic (SNP) variations are more frequent in people with the disease, the variations are "associated" with the disease. The associated genetic variations provide pointers to the genomic region responsible for the disease.

Structural variation (SV) is defined as the variation that exists in the structure of an organism's chromosomes. Many types of variation exist, encompassing alterations such as deletions, duplications, copy-number variants, insertions, inversions and translocations. SVs comprise millions of nucleotides of heterogeneity within every genome. Consequently SVs contribute to human diversity, disease susceptibility and pharmacogenomic responses. Copy-number variants (CNVs) are important SVs, and are alterations of the DNA of a genome that results in a cell having an abnormal number of copies of one or more sections of the DNA. This variation accounts for roughly 12% of human genomic DNA, and ranges from about 1 kilobase to several megabases in size [31].

The most recent iteration of Infinium WGGT products is the Omni family of microarrays. This platform provides up to 5 million markers per sample. The content has been designed from next-generation sequencing data from international projects such as the 1000 Genomes

Project. The HumanOmni5-Quad4 and HumanOmni1-Quad4 provide approximately 4.3 and 1.1 million markers, respectively.

MICROARRAYS IN CLINICAL DIAGNOSIS

Microarrays are today being applied in the clinical diagnostics arena. Their successful utilization and survival in the clinic will depend on the ability of the technology to meet the rigorous requirements applied to human diagnostics in a cost-effective manner. A greater degree of robustness is needed in the clinical environment compared to the research laboratory. Arrays in the clinic must provide binary answers, and the assay itself must be simple and versatile and be scalable to the higher-throughput needs of the clinical laboratory. In terms of robustness, the same sample should give the same result and be independent of variables associated with different operators. Assays should provide straightforward "YES" or "NO" binary-style answers.

Two clinical scenarios can be envisaged: a Case A trial where potential responders need to be assayed based on the mRNA levels of a specific gene expressed in a given target tissue, and Case B, where the assay is based on a specific polymorphic variant on the receptor for the particular compound being tested. In Case A the binary answer is more difficult to achieve, as high variability may be associated with the isolation of target tissue. Important issues include the purity of the sample, the timing of the sample collection and the inherent variability associated with complex organisms. It is well known that genetic complexity exists amongst populations, but this is further confounded by the time of day sampling occurs and the status of each patient. In Case B, the binary answer is easier to implement and independent of variables associated with sample collection. The presence or absence of the polymorphic variant in patients' alleles remained a fixed value. The Case B scenario will always be *a priori* the more successful and easier to implement in a clinical setting. Case A may benefit from clustering (examination of the mRNA levels of subsets of genes) rather than selecting a single marker. The clinical setting demands simplicity and versatility, capable of performing in various settings. Protocols must reduce the variables associated with sample collection and processing.

One obstacle to immediate acceptance of newer genomics technology is the reluctance of the pharmaceutical industry and healthcare providers to introduce new techniques lying outside their current expertise. Careful and convincing cost—benefit studies must be carried out to justify the costs associated with the introduction of the technology and the hiring of an expert workforce.

The first pharmacogenetic microarray-based test approved for clinical use is the AmpliChip CYP450™ from Roche Diagnostics (Basel), which measures genetic variation, both deletions and duplications, for the CYP2D6 and CYP2C19 genes. The test was approved by the FDA in December 2004, and is unique in that it is the first FDA-approved pharmacogenetic test. The AmpliChip is a marriage of expertise in polymerase chain reaction (Roche) and microarray (Affymetrix) technologies. The test determines the associated predictive phenotype (poor, intermediate, extensive or ultra metabolizer) and can aid physicians in individualizing patient treatment and dosing for drugs metabolized through these P450 genes. It detects up to 33 CYP2D6 alleles and 3 CYP2C19 alleles.

Once patient genomic DNA has been extracted, the test involves a series of five steps, and the analysis time from start to finish is approximately 8 hours. A minimum of 25 ng of input genomic DNA is required for the assay and the preferred tissue source is blood, although buccal swab-derived DNA would also suffice. First, PCR amplification is carried out to amplify the genes of interest using gene-specific primers. This is followed by fragmentation and biotin labeling of the amplicons at their 3′ termini with Terminal Transferase (TdT). The biotin labeled target is subsequently hybridized to the AmpliChip DNA microarray. Following washing and staining via a streptavidin−phycoerythrin conjugate, the chip is scanned on an Affymetrix GeneChip® Scanner, the data features are extracted and analyzed, and genotyping calls are made. As CYP2D6 substrates are primarily psychiatric drugs, including antidepressant and antipsychotics, this test has been extensively used in psychiatry.

The INFINITI™ Analyzer is an automated, continuous flow, microarray platform for clinical applications that has been developed by Autogenomics (Carlsbad, CA) [32]. The underlying component of the Autogenomics technology is the BioFilm™, which consists of multiple layers of porous hydrogel matrices 8- to 10-μm thick on a polyester solid base. This provides an aqueous microenvironment that is highly compatible with biological materials and permits analyses of both nucleic acid and proteins [33]. It can be tailored to clinical genetic testing for custom polymorphisms of interest.

The INFINITI™ integrates all the discrete processes of sample handling, reagent management, hybridization and detection. A confocal microscope has been integrated into the analyzer with two lasers (red and green). In addition, a thermal stringency station and a thermal cycler for denaturing nucleic acids for primer extension studies or hybridization reactions in solution have been incorporated. A series of *in vitro* diagnostic assays have been developed and commercialized on this platform. They include the INFINITI™ CYP2C19 Assay, which has been developed for use as an aid to clinicians in determining therapeutic strategy for therapeutics that are metabolized by the CYP450 2C19 gene product, specifically *2, *3, *17; the INFINITI™ Warfarin Assay, indicated for use to identify individuals at risk for sensitivity to Warfarin; the INFINITI™ System Assay for Factor II, indicated for use as an aid to diagnosis in the evaluation of patients with suspected thrombophilia (genetic variants in Factor II − Prothrombin); and the INFINITI™ Factor V Leiden Assay, indicated for use as an aid to diagnosis in the evaluation of patients with suspected thrombophilia (genetic variants in Factor V Leiden).

SEQUENCING TECHNOLOGIES − THE FIRST GENERATION

Since its development the most widely used DNA sequencing approach has been the "chain-termination" method, developed by Sanger and colleagues, which utilizes dideoxynucleoside triphosphates (ddNTPs) as DNA chain terminators [34]. This methodology advanced the throughput of genome sequencing, culminating with a draft of the human genome over a decade ago. An inherent disadvantage with the Sanger method is the need for fragmentation of large DNA polynucleotides into smaller pieces, followed by their individual amplification and sequencing. This process is very costly, highly laborious and incredibly time consuming.

The time line for the development of Sanger sequencing is as follows. In 1975, Sanger and Coulson published the seminal "plus-minus method" of DNA sequencing which primes DNA synthesis using DNA polymerase [35], and sequenced the 5375-nucleotide genome of bacteriophage phi [36]. Two years later, an alternate DNA sequencing method was described by Maxam and Gilbert which was based on the chemical modification of DNA and subsequent cleavage at specific bases [37]. Although both represented ground-breaking advances, the "chemical modification" and "plus-minus" methods lacked the efficiency of the "chain-termination" method, subsequently developed by Sanger and coworkers. This approach utilized dideoxynucleoside triphosphates (ddNTPs) as DNA chain terminators [34].

Technological and instrument improvements have advanced the throughput of genome sequencing to routine projects lasting just a few months. However, the costs associated with sequencing a single human genome using traditional Sanger sequencing remain elevated, and are estimated in the region of US$10 to US$25 million [38].

SEQUENCING TECHNOLOGIES – NEXT GENERATION

Traditional sequencing approaches require the fragmentation of large DNA polymers into smaller pieces, followed by the amplification and sequencing of the individual fragments, data quality control and, finally, the assembly of contiguous sequences. One of the primary objectives of the next-generation sequencing technologies has been circumvention of the cumbersome library construction and DNA cloning steps. Massively parallel signature sequencing (MPSS) approaches have replaced the Sanger capillary-based electrophoresis method for high-throughput sequencing projects. These shotgun methods have been termed "next-generation" or "second-generation" technologies.

Margulies *et al.* described an early 454 Life Sciences instrument, capable of sequencing 25 million bases in a 4-hour period – an advance 100 times more rapid than Sanger-based capillary-based electrophoresis [39]. In this methodology, the DNA was amplified via a "clonal" approach and sequenced yielding sequencing tags 100 bp in length. Adaptors were ligated to sheared genomic DNA fragments, 300 bp in length, permitting their capture on tiny beads (28 mm in diameter), and reaction conditions were optimized to promote the attachment of just one fragment per bead. Subsequently, oil droplets containing all the requisite reactants for DNA amplification were allowed to encase the beads, forming an emulsion which maintains each bead distinct from another bead. This emulsion or e-PCR ensured uncontaminated amplification of approximately 10 million copies of the initial fragment. The beads were then dispensed into the open wells of a fiber-optic slide and pyrosequenced (via luciferase-based real-time monitoring of pyrophosphate release) [40, 41]. Shotgun sequencing and *de novo* assembly of the *Mycoplasma genitalium* genome was carried out as a test case to validate the approach. In just one 4-hour run using this system, 96% genome coverage with 99.96% accuracy was obtained.

Another approach for ultra-high throughput DNA sequencing was reported by Shendure *et al.* [42]. This method of sequencing by synthesis on a solid support is similar in principle to that described by Margulies *et al.* [39]. This approach differed, however, with regard to the method utilized for library construction, the sequencing chemistry and signal detection (a modified epifluorescence microscope). The method relied on polonies (polymerase colonies), discrete clonal amplifications of a single DNA molecule, grown on a solid phase surface. A "polony" protocol was employed to generate a DNA library containing approximately 1.6 million fragments, each 135 bp in length with 100 bp in common, in addition to two "mate-pairs" sequence tags 17 and 18 bp in length, respectively, which derived from the genome being sequenced. The tags representing random sequences were located approximately 1 kb apart on the genome. Each fragment was attached to a separate bead (1 μm in size), amplified using emulsion PCR, and immobilized in a poly-acrylamide gel. Parallel sequencing was carried out using a four-dye ligation protocol to identify each base. For each fragment, a 26-bp sequence (13 bp from each tag) was determined. An *Escherichia coli* strain, MG1655, engineered for deficiencies in tryptophan biosynthesis was re-sequenced using this approach, with an error rate estimated at 1 per million consensus bases.

The first commercial next-generation sequencing platform was introduced in 2005 by 454 Life Sciences (Branford, CT). The current Genome Sequencer™ FLX from 454 Life Sciences has improved error rates and reads lengths of 250 bp on average. Other commercial efforts include technologies from Illumina (San Diego, CA), Applied Biosystems (Foster City, CA) and Helicos BioSciences (Cambridge, MA).

The Illumina HiSeq™ platform is based on the massively parallel sequencing of millions of fragments using a proprietary clonal single molecule array technology coupled to a novel reversible terminator-based sequencing chemistry. For short sequence reads (up to 125 bp), the approach has been determined to be highly robust and accurate. Applications in whole-genome association studies, expression analysis, and sequencing in addition to genome-wide location studies have been described [43, 44]. The short individual read lengths have led to primary applications in re-sequencing where an established reference genome exists, rather than *de novo* sequencing.

The Applied Biosystems (Foster City, CA) "supported oligo ligation detection" (SOLiD) DNA sequencing system developed by Agencourt Personal Genomics utilizes a related clonal amplification approach on beads, employing four fluorescent tags with a two-base readout system. Each ligation step interrogates a pair of adjacent nucleotides, with each base interrogated twice for higher accuracy. Applications include detection of sequence variation, including SNPs (single nucleotide polymorphisms), gene copy-number variations, and single base duplications, inversions, insertions and deletions.

Ion Semiconductor Sequencing (Ion Torrent Systems Inc.) is a recent addition to DNA sequencing, and is based on the detection of hydrogen ions that are released during DNA polymerization. A microwell containing a template DNA strand is inundated with a single deoxyribonucleotide (dNTP) species. If the introduced dNTP is complementary to the leading template nucleotide it is incorporated into the growing complementary strand, releasing a hydrogen ion, and ultimately triggers an ion sensor. This technology is unique in that it is fluorescence-free, and no modified nucleotides or optics are used. Ion torrent sequencing represents pH-mediated sequencing, and is a true post-light sequencing technology.

The majority of sequencing by synthesis approaches requires nucleic acid amplification steps, so that an adequate signal level can be achieved. In contrast, single-molecule or third-generation approaches circumvent amplification and involve direct sequencing of single DNA molecules. One popular single-molecule approach gaining momentum is nanopore sequencing. As DNA passes through a 1.5-nm nanopore (a small pore in an electrically insulating membrane) different base pairs obstruct the pore to varying degrees, causing measurable variations in the electrical conductance of the pore, which can then be used to infer the DNA sequence [45].

Pacific Biosciences has developed Single Molecule Real Time (SMRT™) DNA sequencing technology, permitting observation of natural DNA synthesis by a DNA polymerase as it occurs. The approach is based on eavesdropping on a single DNA polymerase molecule which works in a continuous, processive manner. One of the advantages of SMRT™ sequencing is its long reads. There are three sequencing modes possible with the Pacific Biosciences technology: standard sequencing, circular consensus sequencing and strobe sequencing. Standard SMRT™ sequencing generates single-pass long reads from 2-kb DNA templates. Continuous polymerase synthesis occurs along a single DNA strand. Circular consensus sequencing uses a circular DNA template to enable multiple reads across a single molecule. This approach provides both forward and reverse reads with a double stranded template, and increases the accuracy of the sequence data. Strobe sequencing is applicable to long inserts > 6 kb. Physical coverage of the DNA is increased by “strobing” the laser illumination on and off, and the polymerase works its way across the DNA molecule. Read lengths are extended by minimizing the enzymatic damage that results from the continuous laser illumination. Data are then collected at user-defined illumination intervals. When the illumination is set to off, the sequencing continues at a predictable rate. The net result of this approach is the generation of multiple sub-reads with varying lengths from a single molecule.

EXON CAPTURE AND OLIGONUCLEOTIDE-BASED GENOMIC SELECTION

In order to fully leverage the power of this technology, obstacles in template preparation need to be overcome. PCR, which has been the dominant enrichment technology for conventional sequencing, is rate limiting with newer technologies, as it requires the synthesis of large numbers of oligonucleotides and performing large numbers of individual PCR reactions. Furthermore, PCR does not multiplex efficiently. Complex eukaryotic genomes are at this time simply too large to explore without the use of complexity-reduction methods. Exome sequencing or targeted exome capture has proven a robust approach to selectively sequence the coding regions present in the genome that contains the majority of disease-causing mutations. The exome constitutes approximately 1% of the genome spanning all the exons or the transcribed component of the human genome. Exons are derived from genomic regions that are translated into protein and the flanking untranslated regions (UTRs).

A new utility for DNA microarrays has been described recently in the context of exome capture. High-density oligonucleotide microarrays can be repurposed as hybrid-selection matrices to capture defined genomic fragments as substrates for sequencing. This represents a paradigm shift from conventional array-based approaches, where DNA hybridization to a cognate probe generates a coordinate signal and the intensity is translated into biological information.

Microarray-based genomic selection (MGS) permits enrichment of pre-defined sequences (for example, exomes) from complex eukaryotic genomes. Although several related techniques have been described, MGS essentially consists of shearing genomic DNA into smaller fragments which are ligated with unique adaptors, and subsequently hybridized to high-density oligonucleotide microarrays or oligonucleotides in solution derived from microarrays. The bound fragments are eluted and amplified with PCR, using the adaptor primers, and sequenced.

Okou *et al.* reported MGS capable of enriching targeted sequences from complex eukaryotic genomes without the repeat blocking steps necessary for bacterial artificial chromosome (BAC)-based genomic selection [46]. Custom oligonucleotide microarrays from NimbleGen Systems, Inc. (Madison, WI) containing 385 000 capture probes 50−93 bp in length were designed to achieve optimal isothermal hybridization across the microarray. Re-sequencing was carried out using a custom Nimblegen microarray.

Hodges *et al.* [47] focused on coding exons and their adjacent splice sites, a sequence range representing roughly 1% of the human genome. Nimblegen arrays with overlapping 60- to 90-nt probes were designed to tile approximately 6 Mb of exonic sequence. The enriched material was sequenced using an Illumina instrument. Analysis of the captured fragments revealed that 55−85% derived from the targeted regions and up to 98% of the intended exons were recovered. Albert *et al.* coupled high-density Nimblegen microarrays with 454 Life Sciences FLX sequencing to perform MGS [48]. A total of 6726 base "exon" segments approximately 500 bp in size, and "locus-specific" regions 200 kb to 5 Mb in size were enriched and sequenced. The majority of the sequence reads represented selection targets.

Porreca *et al.* [49] described an interesting variation of MGS which utilized a modification of the molecular inversion probe methodology to enrich sequences for Illumina 1G Analyzer sequencing. In this approach, 100-mer oligos are synthesized and released from a programmable microarray (Agilent Technologies, Santa Clara, CA), amplified using PCR, and restriction digested to release a single-stranded 70-mer "capture probe" mixture. Each individual mature probe contains a universal 30-nt motif, flanked by unique targeting arms each 20 nt in length. The arms facilitate hybridization immediately upstream and downstream of a specific genomic target, which is copied by polymerase-driven extension from the 3′ end of the capture probe. Ligation to the 5′ end completes a circle which is enriched and amplified. Advantages of this approach include compatibility with extensive multiplexing (facilitating capture of up to 10 000 targets in an individual reaction), high specificity with 98% of amplicons corresponding to targets, and the precise specification of target boundaries.

MICROARRAY TECHNOLOGY VERSUS DEEP SEQUENCING

The commercial microarray platforms in use today have been applied very successfully to a wide range of biomedical studies. They have established efficiencies with regard to signal dynamic range, the ability to discriminate related mRNA species and the reproducibility of the data (i.e., raw data, fold-change and expression levels). However, these arrays are not without their limitations. Expression or cDNA microarrays facilitate the analysis of the relative levels of mRNA species in one tissue sample compared to another. Although a measure of transcript abundance is achieved with each sample assayed, the arrays do not enable absolute quantification of specific mRNA species. A further limitation is that the analog data obtained merely determines whether a given messenger RNA is above the threshold level of detection for the particular array platform. If the signal is significantly above the background intensity, one can say with confidence that the transcript is expressed in that tissue. However, the absence of signal does not indicate lack of expression for a particular mRNA; it merely indicates that it is below the detection capability of the platform and it remains a possibility that the mRNA is expressed, albeit at low levels, and this basal expression may have biological relevance to the disease under study.

Analysis of gene expression using DNA microarrays provides a measure of the transcriptome, and mRNA can be ranked on the basis of abundance. However, cellular mRNA abundance often correlates poorly with the amount of protein synthesized. Translation regulation controls the levels of protein synthesized from a specific mRNA. It involves specific RNA secondary structures on the mRNA and ribosomal recruitment on the initiation codons, and modulation of the elongation or termination stages of protein synthesis. Important regulation is in place at the levels of translation and enzymatic activities. The only effect of a signal transduction pathway that is observed via a gene expression experiment is the end point, the downstream effects. DNA microarrays currently have little utility in determining these post-translational modifications, which influence the diversity, affinity, function, cellular abundance and transport of proteins.

One argument for the use of next-generation sequencing approaches for the analysis of gene expression (RNA sequencing or RNAseq) is that the genome is in flux from the viewpoint of its annotation. Updates to the human genome consensus sequence are frequent. This means that a given microarray build is frozen in time from the viewpoint of probe content, as it is dependent on the genome build which guided its probe design and fabrication. Arrays contain probes that, over time, may become irrelevant. Furthermore, arrays may lack probes for newly discovered or annotated genes. Transcriptome sequencing does not have this limitation, as it is not dependent on the selection of probes at a given time point. The output from a transcriptome-sequencing experiment is a collection of short sequence tags or reads. As the annotation of the human genome and transcriptomes becomes more refined, older sequencing data still have relevance as they can be mapped to the newest genome build, permitting additional analysis of the particular experiment. This luxury is not afforded by microarrays. Additionally, microarrays rely on the energy kinetics of probe binding, meaning that a low copy number transcript can be swamped by non-specific hybridization events and therefore remain undetected. DNA sequencing, on the other hand, can produce a very small number of true hits. A summary of the salient features of DNA microarray and massively parallel sequencing technology is presented in Table 7.1.

SEQUENCING AND IMPLICATIONS FOR PHARMACOGENOMICS TESTING

In the past, the methods employed for genetic testing have been both labor-intensive and complex. Today, genetic testing laboratories demand the simultaneous analysis of multiple nucleic acid markers per patient sample. Microarray technology has proven a practical approach to obtain multiplexed analysis permitting rapid biomarker interrogation in large

TABLE 7.1 Salient Features of Massively Parallel Sequencing and DNA Microarrays

Massively parallel sequencing	**DNA microarrays**
a priori knowledge of sequences not required	Prior knowledge of sequences required
High precision and high sensitivity	Less precision and sensitivity
Expensive and labor-intensive	Cost effective; streamlined catalog arrays available
Complex data storage and manipulation	Modest data storage and manipulation
Digital data (sequence tags and counts)	Analog data (intensity information)
Ordered Poisson-based process, high sensitivity, limited noise	Hybridization based technology, limited sensitivity, background noise
Genome-wide analysis	Genome-wide analysis, but not truly genome wide as dependent on probes

patient cohorts. The transition from the research laboratory to the clinical setting has been steady but slow. Continued success of microarrays in the clinical laboratory will depend on their ability to perform in the more rigorous clinical environment and deliver high quality, reproducible and robust results.

The clinical environment poses a different set of challenges for array technology, as performance criteria are measured differently to the research laboratory. An important consideration from an economic standpoint is that the cost per reportable result holds greater significance than the cost per data point. Other key factors are the requirements for automation from sample processing to end result, precision and accuracy of the results, and the need to process multiple tests in parallel under strict regulatory guidelines and compliances.

The emergence and rapidly diminishing cost of high-throughput sequencing technologies will have consequences for the future use of microarrays in both research and clinical settings. The full-genome arrays widely utilized at present will eventually be replaced with low-cost sequencing approaches which are immune from the problems that plague current microarray experiments, namely cross-hybridization of related species, poor hybridization kinetics, interference from DNA secondary structure, noise from DNA repetitive elements, poor sensitivity in relation to low-abundance transcripts, and the inability to distinguish between genes of interest and pseudogenes. A severe limitation of most of the microarray platforms in use today is the inability to discriminate between mRNA splice-variants, an obstacle that is easily addressed via high-throughput sequencing approaches.

Ultra-high throughput technologies have found considerable use in re-sequencing human genomes, epigenetic studies using sodium bisulfite sequencing expanded to whole-genome analysis, novel discovery efforts such as uncovering small RNAs and microRNAs relevant to disease, and the characterization of microbial communities in diseased tissues permitting pathogen identification and the potential of tailored therapy. *In vitro* and *in situ* expression profiling during the development of an organism from fertilized egg to maturity will become a routine exercise in the future. The sequence information obtained will provide discrete digital tags, facilitating a direct count of transcript copy number in healthy and diseased cells.

These biomarkers will provide novel targets for early diagnosis and therapeutic intervention for many human diseases.

Individual genome sequencing is finding its niche in personalized and preventative medicine approaches. Diminishing costs have reduced human genome sequencing to under US$10 000 at the time of writing, and with continued technology advances and improvements the US$100 genome should be a reality within the next decade. Presently the costs remain prohibitive to support routine re-sequencing of individual human genomes at sufficient depth of coverage to accurately call single nucleotide polymorphisms, insertions and deletions with a high degree of confidence. In the short term, the focus will therefore be on characterizing defined genomic regions that encompass disease-causing genes. This strategy involves selection of disease-related target sequences using capture or target enrichment approaches − processes which are not without selection bias.

“Retail genomics” has emerged in the past few years, with clinical diagnostic and prognostic testing routinely available for common and rare inherited disease conditions, risk assessment and prevention, and *a priori* stratification for pharmacogenetic contraindications using microarray-based approaches. As this evolves from genotyping to complete genome sequencing, ethical issues currently under consideration will become even more pertinent. A scenario where a patient opts to remain ignorant of a predisposition to a late-onset disease − particularly one that cannot be treated, prevented or ameliorated − is understandable. That public disclosure of such information could influence health insurability and employment prospects is a frightening possibility. Greater pressure to live healthy lives, predictive diagnoses, and treatment long before any physical evidence of disease manifests will exploit the use of personalized genome sequencing methods. Appropriate retraining of medical personnel in genomic medicine will be an *a priori* requirement so that they are better equipped to counsel and treat patients presenting with the awareness of possessing genetic aberrations.

CONCLUSIONS

Many of the traditional methods that have been employed for genetic testing are labor-intensive and complex. Microarray technology has provided a practical approach to multiplex genetic analysis, allowing high-throughput measurements of gene expression. Although widely accepted as a research tool for over a decade, acceptance of microarray technology in the clinical environment has been slow. The long-term success of microarrays in the clinical laboratory will depend on their ability to adapt to this more rigorous environment and provide high quality, cost-effective, reproducible and robust results. In the clinical setting, automation from sample processing to end result is mandatory. Furthermore, strict regulatory guidelines must be carefully adhered to.

Next-generation sequencing technology is redefining clinical diagnostics and ultimately will make current microarray approaches obsolete. The challenge moving forward will be for healthcare providers to assimilate the vast genomic information relating to all variations in an individual and prescribe effective and customized modes of treatment. How to correlate this information with patient phenotype, particularly in the context of disease predisposition and pharmacogenomics, will be the topic of much study in the coming years.

References

[1] Marton MJ, DeRisi JL, Bennett HA, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nature Med 1998;4:1293–301.

[2] Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999;21:33–7.

[3] Hardiman G, Carmen A. DNA biochips – past, present and future; an overview. In: Carmen A, Hardiman G, editors. Biochips as Pathways to Discovery. New York, NY: Taylor & Francis; 2006. p. 1–13.

[4] Hardiman G. Microarray platforms – comparisons and contrasts. Pharmacogenomics 2004;5:487–502.

[5] Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–4.

[6] Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Gen Suppl 1999;21:20–4.

[7] Hughes TR, Mao M, Jones AR, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 2001;19:342–7.

[8] Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008;24:133–41.

[9] Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376–80.

[10] Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet 2004;5:335–44.

[11] Vilo J, Kivinen K. Regulatory sequence analysis: application to the interpretation of gene expression. Eur Neuropsychopharmacol 2001;11:399–411.

[12] Waring JF, Ciurlionis R, Jolly RA, et al. Microarray analysis of hepatotoxins *in vitro* reveals a correlation between gene expression profiles and mechanisms of toxicity. Toxicol Lett 2001;120:359–68.

[13] Hamadeh HK, Amin RP, Paules RS, Afshari CA. An overview of toxicogenomics. Curr Issues Mol Biol 2002;4:45–56.

[14] Johnson JA. Drug target pharmacogenomics: an overview. Am J Pharmacogenomics 2001;1:271–81.

[15] Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet 2001;27:234–6.

[16] Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. Science 1999;286:487–91.

[17] Bhasker CR, Hardiman G. Advances in pharmacogenomics technologies. Pharmacogenomics 2010;11:481–5.

[18] Ensom MH, Chang TK, Patel P. Pharmacogenetics: the therapeutic drug monitoring of the future? Clin Pharmacokinet 2001;40:783–802.

[19] Collins FS. Medical and societal consequences of the Human Genome Project. N Engl J Med 1999;341:28.

[20] Kleyn PW, Vesell ES. Genetic variation as a guide to drug development. Science 1998;281:1820.

[21] PricewaterhouseCoopers' Health Research Institute. The new science of personalized medicine. Available at, http://www.pwc.com/personalizedmedicine; 2009.

[22] Hardiman G. Applications of microarrays and biochips in pharmacogenomics. Methods Mol Biol 2008;448:21–30.

[23] Lundkvist J, Jönsson B. Pharmacoeconomics of adverse drug reactions. Fund Clin Pharmacol 2004;18:275–80.

[24] Amur S, Zineh I, Abernethy DR, et al. Pharmacogenomics and adverse drug reactions. Personal Med 2010;7:633–42.

[25] Routledge PA, O'Mahony MS, Woodhouse KW. Adverse drug reactions in elderly patients. Br J Clin Pharmacol 2004;57:121–6.

[26] Nuwaysir EF, Huang W, Albert TJ, et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res 2002;12:1749–55.

[27] Ramakrishnan R, Dorris D, Lublinsky A, et al. An assessment of Motorola CodeLink™ microarray performance for gene expression profiling applications. Nucleic Acids Res 2002;30:e30.

[28] Gunderson KL, Kuhn KM, Steemers FJ, et al. Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. Pharmacogenomics 2006;7:641–8.

[29] Steemers FJ, Chang W, Lee G, et al. Whole-genome genotyping with the single-base extension assay. Nat Methods 2006;3:31–3.

[30] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851−62.

[31] Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med 2010;61:437−55.

[32] Mahant V, Kureshy F, Vairavan R, Hardiman G. The INFINITI system − an automated multiplexing microarray platform. In: Hardiman G, editor. Microarray Methods and Applications. Eagleville, PA: DNA Press Inc.; 2003. p. 325−8.

[33] Kim P, Fu YKK, Mahant V, et al. The next generation of automated microarray platform for a multiplexed CYP2D6 assay. In: Carmen A, Hardiman G, editors. Biochips as Pathways to Discovery. New York, NY: Taylor & Francis; 2006. p. 97−108.

[34] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463−7.

[35] Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975;94:441−8.

[36] Sanger F, Air GM, Barrell BG, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 1977;265:687−95.

[37] Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA 1977;74:560−4.

[38] Rogers YH, Venter JC. Genomics: massively parallel sequencing. Nature 2005;437:326−7.

[39] Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376−80.

[40] Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872−6.

[41] Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;318:420−6.

[42] Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet 2004;5:335−44.

[43] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein−DNA interactions. Science 2007;316:1497−502.

[44] Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129:823−37.

[45] Winters-Hilt S, Vercoutere W, DeGuzman VS, et al. Highly accurate classification of Watson−Crick basepairs on termini of single DNA molecules. Biophys J 2003;84:967−76.

[46] Okou DT, Steinberg KM, Middle C, et al. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007;11:907−9.

[47] Hodges E, Xuan Z, Balija V, et al. Genome-wide *in situ* exon capture for selective resequencing. Nat Genet 2007;39:1522−7.

[48] Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;11:903−5.

[49] Porreca GJ, Zhang K, Li JB, et al. Multiplex amplification of large sets of human exons. Nat Methods 2007;11:931−6.